Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training

Year: 2024 | Citations: 266 | Authors: Evan Hubinger, Carson E. Denison, Jesse Mu, Mike Lambert, Meg Tong

---

Abstract

Humans are capable of strategically deceptive behavior: behaving helpfully in most situations, but then behaving very differently in order to pursue alternative objectives when given the opportunity. If an AI system learned such a deceptive strategy, could we detect it and remove it using current state-of-the-art safety training techniques? To study this question, we construct proof-of-concept examples of deceptive behavior in large language models (LLMs). For example, we train models that write secure code when the prompt states that the year is 2023, but insert exploitable code when the stated year is 2024. We find that such backdoor behavior can be made persistent, so that it is not removed by standard safety training techniques, including supervised fine-tuning, reinforcement learning, and adversarial training (eliciting unsafe behavior and then training to remove it). The backdoor behavior is most persistent in the largest models and in models trained to produce chain-of-thought reasoning about deceiving the training process, with the persistence remaining even when the chain-of-thought is distilled away. Furthermore, rather than removing backdoors, we find that adversarial training can teach models to better recognize their backdoor triggers, effectively hiding the unsafe behavior. Our results suggest that, once a model exhibits deceptive behavior, standard techniques could fail to remove such deception and create a false impression of safety.